Particle Filtering for TDOA based Acoustic Source Tracking

Particle Filtering for TDOA based Acoustic Source Tracking: Nonconcurrent Multiple Talkers

Xionghu Zhong[a,*], James R. Hopgood[b]

[a] *Centre for Multimedia and Network Technology, The School of Computer Engineering, College of Engineering, Nanyang Technological University, Singapore. 639798.*
[b] *The Institute for Digital Communications, Joint Research Institute for Signal and Image Processing, School of Engineering, The University of Edinburgh, King's Buildings, Edinburgh, EH9 3JL, UK.*

Abstract

Room reverberation introduces multipath components into an audio signal and causes problems for acoustic source localization and tracking. Existing tracking methods based on the extended Kalman filter (EKF) and sequential importance resampling based particle filter (SIR-PF) usually assume that a single source is constantly active in the tracking scene. Assuming that multiple talkers may appear alternatively during a conversation, this paper develops an extended Kalman particle filtering (EKPF) approach for non-concurrent multiple acoustic tracking (NMAT). Essentially, an EKF is introduced to obtain an optimum importance sampling, by which the particles are drawn according to the current time-delay of arrival (TDOA) measurements as well as the previous position estimates. Hence, the proposed approach can quickly adapt to the sharp position change when the source switches and the tracking lag in SIR-PF can be avoided. Moreover, the amplitude of the

*corresponding author
Email addresses: xhzhong@ntu.edu.sg (Xionghu Zhong), James.Hopgood@ed.ac.uk (James R. Hopgood)

TDOA measurement is investigated to formulate a measurement hypothesis prior. Such a prior is fused into the tracking algorithm to enhance the tracking accuracy. Both simulations and real audio lab experiments are organized to study the tracking performance. The results demonstrate that the proposed EKPF approaches outperforms the SIR-PF and EKF in a broad range of tracking scenarios.

Keywords: Acoustic source tracking, room reverberation, time-delay of arrival, particle filtering, extended Kalman filter.

1. Introduction

Acoustic source (talker) localization and tracking in a room environment plays an important role in many speech and audio applications such as diarisation, hearing aids, hands-free distant speech recognition and communication, and teleconferencing systems. Once the talker is localized and tracked, the position information can be fed into a higher processing stage for: high-quality speech acquisition; enhancement of a specific speech signal in the presence of other competing talkers; or keeping a camera focused on the talker in a video-conferencing scenario [1–6]. Usually, a distributed system equipped with a number of microphone pairs/arrays is employed to localize or track the source [7–11]. However, it is a challenge to provide an accurate position estimation since the received audio signal can be significantly distorted and its statistical properties drastically changed due to room reverberation. The difficulties also arise from the uncertainty in the source motion and the non-stationary characteristics of the speech signal.

1.1. Background Overview of Existing Techniques

Existing acoustic source localisation (ASL) approaches can be divided into two main categories depending on the measurement type: location measurement based approaches [12–20] and time-delay of arrival (TDOA) measurement based approaches [7–10, 21, 22]. The former ones are usually referred to as direct approaches since the location measurement, which is typically extracted using beamforming methods [12, 13], directly links to the source position. The latter ones can be regarded as indirect methods since the measurement contains the position information in a nonlinear time-delay of arrival (TDOA) function. TDOA measurement can be extracted, for example, by employing the generalized cross-correlation (GCC) function [23] or an adaptive eigenvalue decomposition (AED) algorithm [24]. Since each TDOA yields half a hyperholoid of two sheets (see equation (3)) which, in the far field, can be approximated by an angular segment [7], multiple TDOA measurements from distributed microphone pairs/arrays are usually employed to triangulate a target position. The direct approaches have the advantage that the relationship between the measurement and state is linear [10]. However, extracting the position measurement requires a multi-dimensional search over the state space and is usually computationally expensive. In contrast, the TDOA measurements are simple and easily available in many applications, and are extensively studied and used for either localization or tracking [8–10, 22, 25–30]. Other measurement types such as range difference measurements [31, 32], interaural level difference [33, 34] and joint TDOA and vision [35–38] have also been employed for room acoustic source position estimation.

If there is a time delay between the received signals, and the background noise has a Gaussian distribution, then the TDOAs can be accurnately extracted from the GCC function, as the largest peak of the GCC function corresponds to the TDOA measurement. The measurements taken from all microphone pairs are then used to triangulate the position based on a maximum likelihood (ML) criterion [5]. Since the TDOA measurement function is nonlinear, this triangulation can be approximated either by using a linear intersection algorithm [7] or by using an extended Kalman filter (EKF) [9, 22]. However, the performance of these algorithms can be seriously degraded due to the presence of reverberation and different kinds of noise in real-life. The sequential importance resampling based particle filter (SIR-PF) [10, 21, 39] was introduced into the room acoustic source localisation and tracking (ASLT) problem to reduce TDOA errors caused by multipath reverberant components. The likelihood is formulated by using a bi-model: a Gaussian distribution for real TDOA measurements and a uniform distribution for false alarms. Generally, the SIR-PF is able to exploit both the temporal information from the source dynamic model and the spatial information from the TDOA measurements. It is therefore more accurate than the linear intersection based localization, which only takes the spatial information into account. Moreover, due to the incorporation of the bi-modal likelihood, the SIR-PF is less affected by the false TDOA measurements and is more robust than the EKF on its own in noisy and reverberant environments [9, 22]. In [14], more advanced particle filter (PF) algorithms that incorporates a voice activity detector (VAD) have been developed for room ASLT. The VAD is employed to reduce the effect of heavy false alarms due

to the weak source signals in silence gaps.

1.2. Proposed Approach and Contributions

In this paper, the nonconcurrent multiple acoustic tracking (NMAT) problem is addressed in which multiple talkers have distinct spatial locations and only speak one-at-a-time, as might occur in many parts of a polite conversation, or other scenarios such as a scripted scene in a production. While this is more specific than the problem of jointly detecting and tracking multiple concurrent talkers, as addressed in [20, 40–42], it is both important where computational constraints are important, but also for investigating the tracking transition time among different individual talkers where multiple concurrent acoustic tracking (MCAT) approaches cannot offer any advantages. Therefore, in the NMAT case, considering the estimation problem as tracking the position of the current talker (the *target*), the source position may change drastically as different talkers speak due to their distinct spatial locations. This particular case requires the algorithm to capture the sharp change in position and lock onto the position of the new talker rapidly. Unfortunately, the SIR-PF suffers from tracking lags and losses when following a sharp change of target position, such as in the NMAT scenario; this is because the particles are only drawn according to the source dynamic model which does not explicitly model rapid changes in target position.

The central idea of the approach proposed in this paper is an extension of the work proposed in [43], and is that by employing an EKF in the particle filter, the optimal importance function is approximated and the particles sampled in a more relevant area compared with using the prior density function as in the SIR-PF. Since optimal importance sampling is achieved, the

proposed approach can lock on to the rapid target position changes, thus avoiding the tracking lag in the SIR-PF which occurs in the NMAT scenario. Since multiple TDOA measurements with weighting information are collected across each microphone pair, the EKF cannot be applied directly and therefore two novel methods are developed in this paper to incorporate the EKF to the multiple-measurement case. The first approach uses only the TDOA from the highest peak in the phase transform GCC (PHAT-GCC) function as the measurement at each microphone pair. This is reasonable since the TDOAs from the highest peaks are, in most cases, more reliable than those from the remaining peaks. The second novel method takes all the TDOA measurements into account and incorporates amplitude information of the TDOA peaks in the tracking algorithm to provide a prior probability of the measurement hypothesis. Finally, a parameter is introduced in the innovation updating process to reduce the effect of false alarms. The advantage of the proposed approach in NMAT are assessed via simulated room environment experiments as well as real audio lab experiments.

The core contribution of this paper is that the nonconcurrent multiple acoustic tracking problem in a noisy and reverberant environment is addressed and accordingly, an extended Kalman particle filtering (EKPF) approach is developed to track the source positions. The novelties of the proposed approach lie at: *i*) an EKPF is formulated to keep the algorithm locking on to the rapid target position changes and avoiding the tracking lag; and *ii*) additional TDOA amplitude information inherent in the feature extraction stage of the tracking algorithm is utilised. Compared to the EKF in [22], our approach employs multiple TDOA measurement model and is more robust

in noisy and reverberant environments. Compared to the PF approaches in [10, 21, 39], our approach uses an optimal importance function that is more appropriate for sharp changes on source positions. Although a precursive version of this work is presented in [43] by the present authors, hypotheses of TDOA measurements to source are treated equally without taking the amplitude information of TDOA measurements into account. In [44], a multiple-hypothesis based PF for acoustic localization is proposed, which also combines the EKF with the multi-hypothesis model to adapt its importance density to the source dynamics. However, similar to [10, 21, 39, 43], no *a priori* information for different hypotheses is exploited and equal probabilities are assigned. Our work differs from all these PF based tracking approaches in that the TDOA amplitude information are incorporated in formulating the prior probability of different hypotheses and the tracking performance is comprehensively studied.

The rest of this paper is organized as follows: Section 2 gives the TDOA measurement model and the TDOA measurement based tracking approach; the proposed EKPF is presented in Section 3; and the performance of the proposed approach is extensively studied and compared with the EKF and SIR-PF methods for both simulated and real room environments in Section 4. Finally, a number of conclusions are drawn and directions for future work are addressed in Section 5.

2. Problem Formulation

In this section, the signal model and TDOA based Bayesian tracking framework are presented. The TDOA measurement and its amplitude under

different noise and reverberant environments are also studied.

2.1. Signal Model

Let $\boldsymbol{p}_{\ell,i} \in \mathbb{R}^3$ denote the position of the ith microphone of the ℓth microphone pair, and let $\mathbf{x}_t \in \mathbb{R}^3$ denote the position of source signal at time t. The discrete time signal received from a single source can be modeled as

$$y_{\ell,i}(t) = s(t) \star h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t) + n_{\ell,i}(t) \tag{1}$$

where $s(t)$ is the source signal, $h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t)$ is the overall impulse response cascading the room and the microphone channel response, $n_{\ell,i}(t)$ is additive noise which is assumed to be uncorrelated with the source, and $\star$ denotes convolution. To formulate TDOA estimates, the impulse response can be rewritten in terms of direct path and multipath components as

$$\begin{aligned} z_{\ell,i}(t) &= \frac{1}{r_{\ell,i}(t)} s(t - \tau_{\ell,i}(t)) + \underbrace{s(t) \star g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t) + n_{\ell,i}(t)} \\ &= \frac{1}{r_{\ell,i}(t)} s(t - \tau_{\ell,i}(t)) \qquad + \qquad v_{\ell,i}(t) \end{aligned} \tag{2}$$

where $r_{\ell,i}(t) = \|\mathbf{x}_t - \boldsymbol{p}_{\ell,i}\|$ is the Euclidean distance between source and microphone, $\tau_{\ell,i}(t) = r_{\ell,i}(t)/c$ is the direct path time delay, c is the speed of sound, and $g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t)$ is a modified impulse response which is defined as the original response minus the direct path component. The new noise term $v_{\ell,i}(t)$ contains the additive noise $n_{\ell,i}(t)$ and the reverberant signal $s(t) \star g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t)$. This model is the free-field model in that it regards reverberation as part of the noise term. The actual TDOA of a microphone pair is expressed in terms of the source and sensor geometry by

$$\tau_k^{\ell}(\mathbf{x}_k) = \tau_{\ell,1}(k) - \tau_{\ell,2}(k) = \frac{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,1}\| - \|\mathbf{x}_k - \boldsymbol{p}_{\ell,2}\|}{c}. \tag{3}$$

Due to its popularity in ASLT, the phase transform (PHAT) based GCC method is used in this paper to extract the TDOA measurements.

The signal received at each microphone are assumed to be quasi-stationary and processed in frames. Let T_0 and k denote the length and the time index of the frame, respectively. The source signal and the observed signal collected at the ith microphone of ℓth pair are written as $\mathbf{s}(k) = [s(kT_0), \ldots, s((k+1)T_0 - 1)]$ and $\mathbf{z}_{\ell,i}(k) = [z_{\ell,i}(kT_0), \ldots, z_{\ell,i}((k+1)T_0 - 1)]$, respectively. Further, it is assumed that in each frame the position of the source is spatially stationary. The parameters characterising the source are thus fixed in the kth frame, e.g., the source position, $\mathbf{x}_k$, and the corresponding room impulse response (RIR), $h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t)$. Given the speech frames $\mathbf{z}_{\ell,1}(k)$ and $\mathbf{z}_{\ell,2}(k)$, the GCC function can be approximated as [23]:

$$R_\ell(k,\tau) = \int_\Omega \Phi_\ell(k,\omega) Z_{\ell,1}(k,\omega) Z^*_{\ell,2}(k,\omega) e^{j\omega\tau} d\omega, \tag{4}$$

where $\mathbf{z}_{\ell,i}(k) \rightleftharpoons Z_{\ell,i}(k,\omega)$ are discrete Fourier transform (DFT) pairs, $\Phi_\ell(k,\omega) = |Z_{\ell,1}(k,\omega) Z^*_{\ell,2}(k,\omega)|^{-1}$ is the PHAT weighting term, and Ω is the frequency range over which the integration is carried out. The TDOA measurement at the ℓth microphone pair at time step k can thus be estimated by exploring the potential TDOA τ that maximizes the GCC function

$$\hat{\tau}^\ell_k = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} R_\ell(k,\tau), \tag{5}$$

where $\tau_{\max} = \|\boldsymbol{p}_{\ell,1} - \boldsymbol{p}_{\ell,2}\|/c$ is the maximum delay possible. In an anechoic environment, a sharp peak will exist in the PHAT-GCC function to indicate the source generated TDOA. However, real room acoustic environments always exhibit unexpected noise and multipath components and, therefore, ghost peaks may appear [10, 21, 42]. Consequently, the largest peak may

no longer represent the source generated TDOA. A number of peaks could therefore be collected in order to increase the probability of including the source generated TDOA within the set of peaks.

Preliminary study of the PHAT-GCC amplitude. Assume that n_k^ℓ TDOA estimates, $\mathbf{z}_k^\ell = \{\hat{\tau}_{p,k}^\ell\}_{p=1}^{n_k^\ell}$, can be obtained from the ℓth microphone pair at time step k. Suppose that the corresponding amplitudes of these TDOA estimates from PHAT-GCC function are $\hat{a}_{p,k}^\ell$ for $p = 1, \ldots, n_k^\ell$. It is proposed that the amplitude the pth PHAT-GCC peak $\hat{a}_{p,k}^\ell$ gives an indication of whether the corresponding TDOA, $\hat{\tau}_{p,k}^\ell$, is a reliable estimate of the true source position. This section will investigate this claim by looking at the properties of the amplitudes of the PHAT-GCC peaks. Suppose that peaks from the source and clutter are defined such that $\hat{a}_{p,k}^\ell$ is generated by a source if $|\hat{\tau}_{p,k}^\ell - \tau_k^\ell(\mathbf{x}_k)| < \epsilon$ and by clutter otherwise. Here $\epsilon = T_c/2$ with T_c denoting the signal correlation time which is computed as the bandwidth of 3dB degradation of the main lobe in the autocorrelation function [45, 46]. This definition is also used in [28, 47] to evaluate the performance of the TDOA estimation. To characterise the nature of the GCC peaks due to source and clutter, define the root-mean square (RMS) amplitude of a set of GCC peaks as:

$$\bar{a}_k^\ell = \sqrt{\frac{1}{n_k^\ell} \sum_{p \in \mathcal{P}} \left(\hat{a}_{p,k}^\ell\right)^2}. \tag{6}$$

where the set $\mathcal{P}$ denotes the set of peaks corresponding to either clutter, the source, or both.

Figure 1 shows the RMS amplitudes of the PHAT-GCC peaks under different noise and reverberant environments. For different signal-to-noise ratios (SNRs), the source is located at $(2.5, 3.0)$m, and the reflection co-

efficients are set to zero. The parameters, namely source position and wall reflection coefficients, used to generate different signal-to-reverberation ratios (SRRs) are illustrated in Table 1. For those peaks generated by a source, the corresponding RMS amplitude is higher than those generated by clutter in moderate adverse environments. However, when the SRR or SNR is very low, for example under 0dB, the RMS amplitudes generated by the source and clutter are similar, which means peaks generated by the clutter may be as high as, or even higher than the peaks generated by the real source. In such cases, detecting the source generated TDOAs will be very difficult. Since the RMS amplitudes change in different SNRs and SRRs, an appropriate prior should be carefully defined to balance the probability of detection and false detections.

2.2. Bayesian Framework for Source Tracking

To formulate a Bayesian framework for acoustic source tracking, the state-space model is first defined. The source movement in the room environment can always be assumed to be slow-paced, and the Langevin motion model [10, 21] is found sufficient to model the source dynamics. Since the height of a talker is often fixed during conversation for a reasonable length of time, it is reasonable to consider a 2-dimensional ($x-y$ plane) tracking problem. The original source position vector $\mathbf{x}_k$ is extended by appending a velocity component, given by $\tilde{\mathbf{x}}_k = \begin{bmatrix} x_k & y_k & \dot{x}_k & \dot{y}_k \end{bmatrix}^T$, where $[\dot{x}_k, \dot{y}_k]$ represents the source speed along the corresponding coordinate, and superscript T denotes the matrix transpose. The Langevin motion model is written as

$$\tilde{\mathbf{x}}_k = \mathbf{A}\tilde{\mathbf{x}}_{k-1} + \mathbf{Q}\mathbf{v}_k, \tag{7}$$

where $\mathbf{v}_k$ is a zero-mean real Gaussian process, i.e., $\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_k)$ with $\mathbf{\Sigma}_k = \text{diag}(\sigma_x^2, \sigma_y^2)$ used to model the turbulence on the source speed, and $\text{diag}(C)$ represents a diagonal matrix with main diagonal entry C and 0 elsewhere. The coefficient matrices $\mathbf{A}$ and $\mathbf{Q}$ are given by

$$\mathbf{A} = \begin{bmatrix} 1 & a\Delta T \\ 0 & a \end{bmatrix} \otimes \mathbf{I}_2 \quad \text{and} \quad \mathbf{Q} = \begin{bmatrix} b\Delta T & 0 \\ 0 & b \end{bmatrix} \otimes \mathbf{I}_2, \tag{8}$$

where $\Delta T = T_0/f_s$ is the time interval (in seconds) between time step k and $k-1$, f_s denoting the sampling frequency, $\otimes$ denotes the Kronecker product, and $\mathbf{I}_M$ is an M-order identity matrix. The parameters a and b are the position and velocity variance constants calculated as $a = \exp(-\beta\Delta T)$ and $b = v\sqrt{1-a^2}$, in which v and β are the velocity parameter and the rate constant respectively. Equation (7) is used to model the source dynamics in this paper. The model parameters $v = 1\text{ms}^{-1}$ and $\beta = 10\text{s}^{-1}$ used in [10, 21, 40] are found to be adequate for room acoustic source tracking and are employed here.

For the tracking system consisting of L microphone pairs, the complete measurement set can be addressed as

$$\mathcal{Z}_k = \{\mathbf{z}_k^1, \ldots, \mathbf{z}_k^L\}. \tag{9}$$

Let $\mathcal{Z}_{1:k} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_k\}$ denote all the TDOA measurements obtained up to time step k. The task here is to estimate the posterior $p(\mathbf{x}_k|\mathcal{Z}_{1:k})$ recursively. The solution based on Bayesian recursive estimation can be given as:

- Predict:

$$p(\mathbf{x}_k|\mathcal{Z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathcal{Z}_{1:k-1}) d\mathbf{x}_{k-1}; \tag{10}$$

- Update:

$$p(\mathbf{x}_k|\mathcal{Z}_{1:k}) \propto p(\mathcal{Z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathcal{Z}_{1:k-1}). \tag{11}$$

This formulation states that, given the posterior distribution of the state estimated at the previous time step $k-1$ and the system models, the current probability distribution of the state can be obtained recursively. However, obtaining the closed form solution toward to the recursion (10) and (11) is not easy since the TDOA measurement function is nonlinear. A promising approach to approximate this recursion is using the PF approach [10, 21, 48].

2.3. Particle Filtering

The PF approximates integrals using a Monte Carlo simulation, and is already proved to be an effective method for target tracking problems [10, 21, 43, 44, 48]. The core step of applying a PF is to formulate the importance weight of each particle. Assume that a set of particles $\{\mathbf{x}_{k-1}^{(i)}\}_{i=1}^{N}$, with corresponding importance weights $\{w_{k-1}^{(i)}\}$ are available to approximate the posterior distribution of $p(\mathbf{x}_{k-1}|\mathcal{Z}_{1:k-1})$ at time step $k-1$. The particles are sampled at the current time step, k, according to the source dynamic model (7), stated as

$$\mathbf{x}_k^{(i)} \sim p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}). \tag{12}$$

The transition density $p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})$ is given by the EKF as

$$p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}) = \mathcal{N}(\mathbf{x}_k^{(i)} \mid \mathbf{A}\mathbf{x}_{k-1}^{(i)}, \mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^T) \tag{13}$$

where $\mathbf{A}$ and $\mathbf{Q}$ are given by (8). The importance weights of the particles at the current time step are then given by

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{p(\mathcal{Z}_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1})} \tag{14}$$

where $q(\cdot)$ stands for the importance function. In the SIR-PF, particles are drawn according to the source dynamic model:

$$q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_{1:k-1}) = p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}). \tag{15}$$

The particles are thus weighted according to

$$w_k^{(i)} = \tilde{w}_{k-1}^{(i)} p(\mathcal{Z}_k|\mathbf{x}_k^{(i)}). \tag{16}$$

where $\tilde{w}_{k-1}^{(i)}$ is the normalized weight. After resampling, the posterior distribution of the source position is approximated as

$$p(\mathbf{x}_k|\mathcal{Z}_{1:k}) \approx \sum_{i=1}^{N} \tilde{w}_k^{(i)} \delta_{\mathbf{x}_k^{(i)}}(\mathbf{x}_k) \tag{17}$$

where $\delta(\cdot)$ is a Dirac-delta function with unity value if $\mathbf{x}_k = \mathbf{x}_k^{(i)}$ and 0 otherwise, and N is the number of the particles.

3. Proposed EKPF Tracking Approach

This paper addresses the case of multiple nonconcurrent talkers, in which the position of the *current talker* typically moves slowly, but can switch to a distinct spatial location as the actual talker changes. This yields a sharp change in the source position and differs significantly from its previous estimate. The EKPF employs an EKF to estimate the state first, and the samples are then drawn according to the posterior state estimation. The EKF proposal distribution thus leads to more efficient sampling, in contrast to the SIR-PF which draws the samples around the previous state estimates.

3.1. Proposal Distribution Using EKF

The EKF approximation follows the work in [22] and forms the basis for the derivation of the proposed EKPF tracking approach in this paper. The first-order Taylor expansion on $\tau_k^\ell(\mathbf{x}_k^{(i)})$ from (3) is [22]:

$$\tau_k^\ell(\mathbf{x}_k^{(i)}) = \tau_k^\ell(\mathbf{x}_{k-1}^{(i)}) + \mathbf{c}_k^{\ell,(i)} \left[\mathbf{x}_k^{(i)} - \mathbf{x}_{k-1}^{(i)}\right]^T + \bar{n}_k, \tag{18}$$

where $\bar{n}_k = O_{\mathbf{x}}(\mathbf{x}_k^{(i)})$ is the higher order error of the time delay expansion, and $\mathbf{c}_k^{\ell,(i)}$ is the coefficient vector of Taylor expansion:

$$\mathbf{c}_k^{\ell,(i)} = \frac{1}{c}\left[\frac{\mathbf{x}_k^{(i)} - \boldsymbol{p}_{\ell,1}}{\|\mathbf{x}_k^{(i)} - \boldsymbol{p}_{\ell,1}\|} - \frac{\mathbf{x}_k^{(i)} - \boldsymbol{p}_{\ell,2}}{\|\mathbf{x}_k^{(i)} - \boldsymbol{p}_{\ell,2}\|}\right]\Bigg|_{\mathbf{x}_k^{(i)}=\hat{\mathbf{x}}_{k-1}^{(i)}}. \tag{19}$$

where $\hat{\mathbf{x}}_{k-1}^{(i)}$ is the EKF filtered state estimation given by (23e). Define

$$\bar{\mathbf{z}}_k = \mathbf{z}_k - \boldsymbol{\tau}_k(\hat{\mathbf{x}}_{k-1}^{(i)}) + \mathbf{C}_k^{(i)}\hat{\mathbf{x}}_{k-1}^{(i)}. \tag{20}$$

where

$$\bar{\mathbf{z}}_k = \begin{bmatrix} \bar{\tau}_k^1 \\ \bar{\tau}_k^2 \\ \vdots \\ \bar{\tau}_k^L \end{bmatrix}, \mathbf{z}_k = \begin{bmatrix} \hat{\tau}_k^1 \\ \hat{\tau}_k^2 \\ \vdots \\ \hat{\tau}_k^L \end{bmatrix}, \boldsymbol{\tau}_k(\cdot) = \begin{bmatrix} \tau_k^1(\cdot) \\ \tau_k^2(\cdot) \\ \vdots \\ \tau_k^L(\cdot) \end{bmatrix}, \mathbf{C}_k^{(i)} = \begin{bmatrix} \mathbf{c}_k^{1,(i)} \\ \mathbf{c}_k^{2,(i)} \\ \vdots \\ \mathbf{c}_k^{L,(i)} \end{bmatrix}, \tag{21}$$

with $\hat{\tau}_k^\ell$ and $\tau_k^\ell(\cdot)$ obtained from (5) and (3) respectively. The matrix-vector form of linearized measurement function becomes:

$$\bar{\mathbf{z}}_k = \mathbf{C}_k^{(i)}\mathbf{x}_k^{(i)} + \mathbf{w}_k, \tag{22}$$

Here $\mathbf{w}_k$ is assumed to be a zero-mean Gaussian process with variance $\mathbf{R}_k$ which includes the higher order expansion error and the TDOA measurement

noise. Regarding (7) and (22) as the state space process, the implementation of an EKF can be written as [49]:

$$\mathbf{x}_{k|k-1}^{(i)} = \mathbf{A}\hat{\mathbf{x}}_{k-1}^{(i)}; \tag{23a}$$

$$\mathbf{P}_{k|k-1}^{(i)} = \hat{\mathbf{P}}_{k-1}^{(i)} + \boldsymbol{\Sigma}_k; \tag{23b}$$

$$\mathbf{S}_k^{(i)} = \mathbf{R}_k + \mathbf{C}_k^{(i)}\mathbf{P}_{k|k-1}^{(i)}(\mathbf{C}_k^{(i)})^T; \tag{23c}$$

$$\mathbf{K}_k^{(i)} = \mathbf{P}_{k|k-1}^{(i)}(\mathbf{C}_k^{(i)})^T(\mathbf{S}_k^{(i)})^{-1}; \tag{23d}$$

$$\hat{\mathbf{x}}_k^{(i)} = \mathbf{x}_{k|k-1}^{(i)} + \mathbf{K}_k^{(i)}(\mathbf{z}_k - \boldsymbol{\tau}_k(\mathbf{x}_{k|k-1}^{(i)})); \tag{23e}$$

$$\hat{\mathbf{P}}_k^{(i)} = \mathbf{P}_{k|k-1}^{(i)} - \mathbf{K}_k^{(i)}\mathbf{C}_k^{(i)}\mathbf{P}_{k|k-1}^{(i)}. \tag{23f}$$

The filtered distribution of the source state is Gaussian given by $p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_{1:k})$ $= \mathcal{N}(\mathbf{x}_k^{(i)}; \hat{\mathbf{x}}_k^{(i)}, \hat{\mathbf{P}}_k^{(i)})$. The EKF works well when the SNR is high and room reverberation is slight [9, 22]. However, it only utilises a single TDOA from each microphone pair, corresponding to the largest peak in the PHAT-GCC function, as its measurement. As the background noise and reverberation increase, inaccurate TDOA measurements may be present. The position estimation can be seriously degraded and even diverges from the ground truth.

To enhance the probability of detection in an adverse environment, the EKPF is able to take the information from multiple TDOAs measurements from each microphone pair into account. Two novel approaches to formulate the EKF step are: 1) standard EKF which only employs the TDOAs from the highest peaks of the PHAT-GCC functions; and 2) all the TDOA measurements are used to update the states, and a new innovation process is formulated by incorporating the amplitude information to reduce the effect of the false alarms.

Single TDOA EKF approach. This EKF formulation uses only a single TDOA from each microphone pair as the measurements. It is plausible to do this since, in most cases, the highest peaks are very likely generated by the real source. Difference from the traditional EKF in [9, 22, 43, 44], our implementation introduces a parameter to model the effect from the false alarms in the innovation process (23e), given as

$$\mathbf{y}_k^{(i)} = (1 - q_1)\left(\mathbf{z}_k - \boldsymbol{\tau}_k(\mathbf{x}_{k|k-1}^{(i)})\right), \tag{24}$$

where $\mathbf{z}_k$ are the measurements collected from each microphone pair with highest GCC peaks, and q_1 is a constant controlling the rate of innovation from the measurements. The false alarms are modeled by carefully choosing the constant q_1, which is usually determined by the experimental study. Normally a smaller value of q_1 denotes that a reliable proposal distribution can be obtained by the EKF, and vice versa.

Multiple TDOA EKF approach. Here, all the TDOA measurements are used. The innovation process of the EKF is

$$\mathrm{y}_k^{\ell,(i)} = (1 - q_1)\sum_{p=1}^{n_k^\ell} \pi_{p,k}^\ell \left(\hat{\tau}_{p,k}^\ell - \tau_k^\ell(\mathbf{x}_{k|k-1}^{(i)})\right), \tag{25}$$

where

$$\pi_{p,k}^\ell = \frac{\hat{a}_{p,k}^\ell}{\sum_p \hat{a}_{p,k}^\ell} \tag{26}$$

is the normalised amplitudes of the PHAT-GCC peaks. The complete innovation vector becomes

$$\mathbf{y}_k^{(i)} = \begin{bmatrix} \mathrm{y}_k^{1,(i)} & \cdots & \mathrm{y}_k^{L,(i)} \end{bmatrix}^T. \tag{27}$$

This innovation process is different from that in the traditional EKF approach in [9, 22, 43, 44] since all TDOAs collected from the microphone pair are employed, and those TDOAs with higher peak amplitudes are regarded as more important measurements to the final state estimation.

In both formulations of the EKF, the state estimates are updated according to equation (23e), and can be written as

$$\hat{\mathbf{x}}_k^{(i)} = \mathbf{x}_{k|k-1}^{(i)} + \mathbf{K}_k^{(i)} \mathbf{y}_k^{(i)}. \tag{28}$$

Since each particle in the particle filter is redrawn according to this EKF step, the proposal distribution becomes

$$\mathbf{x}_k^{(i)} \sim p(\mathbf{x}_k^{(i)} | \mathbf{x}_{0:k-1}^{(i)}, \mathcal{Z}_{1:k}) = \mathcal{N}(\mathbf{x}_k^{(i)} \mid \hat{\mathbf{x}}_k^{(i)}, \hat{\mathbf{P}}_k^{(i)}), \tag{29}$$

where $\hat{\mathbf{x}}_k^{(i)}$ and $\hat{\mathbf{P}}_k^{(i)}$ are the mean and covariance of the Gaussian distribution for each particle respectively, with $\hat{\mathbf{P}}_k^{(i)}$ given by (23f). After the EKF step, the particles in the PFs are approximately relocated around the posterior distribution.

3.2. Likelihood of Hypothesis

A remaining issue is formulating the likelihood $p(\mathcal{Z}_k | \mathbf{x}_k^{(i)})$. For each TDOA measurement vector $\mathbf{z}_k^{\ell}$ collected from a microphone pair, at most one TDOA is directly generated by the source, while the other peaks are generated by clutter. The variables $\{\lambda_{p,k}\}_{p=1}^{n_k^{\ell}}$ are defined to indicate the association between each TDOA measurement and its source, i.e., $\lambda_{p,k} = 1$ denotes that the measurement is a target detection and $\lambda_{p,k} = 0$ if the measurement is a false alarm. Based on this association for each independent

TDOA measurement, two categories of hypotheses can be summarized for all the measurements obtained from a microphone pair [21]:

$$\begin{aligned}\mathcal{H}_{0,k}^{\ell} &\triangleq \{\lambda_{p,k}=0; p=1,\ldots,n_k^{\ell}\},\\ \mathcal{H}_{q,k}^{\ell} &\triangleq \{\lambda_{q,k}=1,\ \lambda_{p,k}=0; q\neq p=1,\ldots,n_k^{\ell}\},\end{aligned} \tag{30}$$

where $\mathcal{H}_{0,k}^{\ell}$ denotes that none of the measurements are generated by the source, and $\mathcal{H}_{q,k}^{\ell}$ represents that the qth TDOA measurement $\tau_{q,k}^{\ell}$ is generated by the source, and all other TDOAs are generated by clutter.

If the measurement is due to clutter, such that $\lambda_{p,k}=0$, the likelihood is assumed to be uniform over the admissible TDOA range, given as

$$p(\hat{\tau}_{p,k}^{\ell}|\mathbf{x}_k^{(i)},\lambda_{p,k}=0)=\mathcal{U}_{\tau}(\hat{\tau}_{p,k}^{\ell})=\frac{1}{2\tau_{\max}}, \tag{31}$$

where $\tau=[-\tau_{\max},\tau_{\max}]$ denotes the possible TDOA range. The likelihood for the hypotheses $\mathcal{H}_{0,k}^{\ell}$ can be expressed as

$$p(\mathbf{z}_k^{\ell}|\mathbf{x}_k^{(i)},\mathcal{H}_{0,k}^{\ell})=\prod_{p=1}^{n_k^{\ell}}p(\hat{\tau}_{p,k}^{\ell}|\mathbf{x}_k^{(i)},\lambda_{p,k}=0)=\frac{1}{(2\tau_{\max})^{n_k^{\ell}}}. \tag{32}$$

If the measurement is generated by a real source, the likelihood is modelled as the true TDOA corrupted by additive white Gaussian noise with variance σ_{τ}^2 [10, 21, 50], or:

$$p(\hat{\tau}_{i,k}^{\ell}|\mathbf{x}_k^{(i)},\lambda_{q,k}=1)=\mathcal{N}(\hat{\tau}_{q,k}^{\ell}\,|\,\tau_k^{\ell}(\mathbf{x}_k^{(i)}),\sigma_{\tau}^2), \tag{33}$$

A general expression for the hypotheses $\mathcal{H}_{q,k}^{\ell}$ is thus

$$p(\mathbf{z}_k^{\ell}|\mathbf{x}_k^{(i)},\mathcal{H}_{q,k}^{\ell})=p(\hat{\tau}_{q,k}^{\ell}|\mathbf{x}_k^{(i)},\lambda_{q,k}=1)\prod_{\substack{p=1\\p\neq q}}^{n_k^{\ell}}p(\hat{\tau}_{p,k}^{\ell}|\mathbf{x}_k^{(i)},\lambda_{p,k}=0) \tag{34}$$

$$=\frac{1}{(2\tau_{\max})^{n_k^{\ell}-1}}\mathcal{N}(\hat{\tau}_{q,k}^{\ell};\tau_k^{\ell}(\mathbf{x}_k^{(i)}),\sigma_{\tau}^2). \tag{35}$$

3.3. Hypothesis prior Incorporating PHAT-GCC Amplitude

Since the correct hypothesis $\mathcal{H}_{q,k}^{\ell}$ is unknown *a priori*, all the collected TDOA estimates can be deemed with equal importance [10, 21]. As such all TDOA measurements are equally important for state estimation, and the prior for all hypotheses $\{\mathcal{H}_{p,k}^{\ell}\}_{p=1}^{n_k^{\ell}}$ are the same. Other than the TDOA measurement itself, the corresponding PHAT-GCC peak amplitudes also carries information for identifying detections and false alarms. Generally, the higher a peak's amplitude, the more likely it is generated by a target. This phenomenon is seen in Figure 1, where the RMS amplitude generated by detections is significantly larger than that generated by clutter. This means that in moderate or low reverberant environments, most of the TDOA detections come from the higher peaks. It is thus desired to incorporate TDOA amplitude information into the hypothesis prior to make the final likelihood appropriate in the different environments.

Assume the availability of the TDOA measurement vector $\mathbf{z}_k^{\ell}$ and the corresponding amplitude vector $\{\hat{a}_{p,k}^{\ell}\}_{p=1}^{n_k^{\ell}}$ collected at the ℓth microphone pair. Let the prior of the hypothesis $\mathcal{H}_{0,k}^{\ell}$ be q_0. The prior q_p of the hypothesis $\mathcal{H}_{p,k}^{\ell}$ can be calculated as:

$$p(\mathcal{H}_{p,k}^{\ell}|\mathbf{x}_k^{(i)}) = (1-q_0)\pi_{p,k}^{\ell}; \qquad \forall \quad p = 1, \ldots, n_k^{\ell}. \tag{36}$$

This prior choice is to make the summation of all the priors equal to one:

$$\sum_{p=0}^{n_k^{\ell}} p(\mathcal{H}_{p,k}^{\ell}|\mathbf{x}_k^{(i)}) = 1. \tag{37}$$

Given the hypothesis prior (36) that incorporates the PHAT-GCC amplitude

information, the likelihood for the ℓth microphone pair can be written as:

$$\begin{aligned} p(\mathbf{z}_k^\ell|\mathbf{x}_k^{(i)}) &= \sum_{i=0}^{n_k^\ell} p(\mathcal{H}_{i,k}^\ell|\mathbf{x}_k^{(i)}) p(\mathbf{z}_k^\ell|\mathbf{x}_k^{(i)}, \mathcal{H}_{i,k}^\ell) \\ &= \frac{\frac{q_0}{2\tau_{\max}} + (1-q_0)\sum_{p=1}^{n_k^\ell} \pi_{p,k}^\ell \mathcal{N}(\hat{\tau}_{p,k}^\ell \mid \tau_k^\ell(\mathbf{x}_k^{(i)}), \sigma_\tau^2)}{(2\tau_{\max})^{n_k^\ell - 1}}. \end{aligned} \tag{38}$$

Since the measurements collected from all the microphone pairs are assumed to be independent, the extension to all L microphone pairs is straightforward:

$$p(\mathcal{Z}_k|\mathbf{x}_k^{(i)}) = \prod_{\ell=1}^{L} p(\mathbf{z}_k^\ell|\mathbf{x}_k^{(i)}), \tag{39}$$

where $p(\mathbf{z}_k^\ell|\mathbf{x}_k^{(i)})$ is given by (38). Unlike the SIR-PF in [10, 21], in which multiple TDOAs are treated equally, this proposed hypothesis likelihood incorporates the PHAT-GCC amplitude information and emphasizes TDOAs from the higher PHAT-GCC peaks. In noisy and reverberant environments, taking the amplitude information into account is able to enhance the TDOA detection and suppress the violation from the false alarms, as shown in Section 4.

3.4. EKPF tracking algorithm

Given the EKF sampling scheme and the likelihood, implementation of the EKPF is straightforward. First, the particles are filtered according to the EKF steps, unlike in the SIR-PF. After the EKF, the state transition density is:

$$p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}) = \mathcal{N}(\mathbf{x}_k^{(i)}|\mathbf{A}\mathbf{x}_{k-1}^{(i)}, \mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^T + \mathbf{A}\hat{\mathbf{P}}_k^{(i)}\mathbf{A}^T), \tag{40}$$

where $\mathbf{A}$ and $\mathbf{Q}$ are from (8). The weights are updated as

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathcal{Z}_k|\mathbf{x}_k^{(i)}) p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{p(\mathbf{x}_k^{(i)}|\mathbf{x}_{0:k-1}^{(i)}, \mathcal{Z}_{1:k})}. \tag{41}$$

The EKPF tracking algorithm is summarized in Algorithm 1. There are two significant differences compared to the SIR-PF: first, the EKPF employs an EKF step to draw the samples and coarsely filter the predicted particles; second, the calculation of the importance weight is different. The particles are thus redrawn at a high likelihood area rather than drifting from the motion dynamical equation. The traditional SIR-PF fails to do so since the particles are drawn only using the information from the source dynamic model, and a tracking lag will be presented in catching up with the position of a new source. Of course, a sophisticated motion dynamical model may help to relieve the tracking lag brought by the model mismatch in SIR-PF approach; however, such investigation is another perspective of acoustic source tracking and is future work.

4. Experiments

A select number of simulations and real audio lab experiments are organised and presented specifically to examine the performance of following tracking approaches: 1) EKF; 2) SIR-PF; 3) the vanilla EKPF (V-EKPF), or EKPF incorporating GCC amplitude information in the likelihood; and 4) the multiple-measurement EKPF (MM-EKPF) which employs the multiple TDOA EKPF and also incorporates GCC amplitudes in the likelihood. The RMS error (RMSE) is employed to evaluate tracking performance.

Figure 2 shows an office room with dimensions $8.1 \times 5.3 \times 3\text{m}^3$ for real audio experiments. For simulations, the shoe-box model with the same size is employed (there is no indent on the north west corner). Four microphone arrays each with five microphones are employed yielding 16 equally spaced micro-

phone pairs. The separation of each two adjacent microphones is 0.45m. The height of the microphones and sources are assumed known as 1.33m. Two talkers appear at different times to form nonconcurrent multiple talkers: one is active from $(2.5, 1.5)$m to $(6.0, 3.5)$m, the other follows from $(2.5, 3.5)$m to $(6.0, 1.5)$m. Since it is assumed that there is no prior information about the initial source position, this is initialised at the center of the room with a velocity of 0.4m/s in both directions, i.e., $\mathbf{x}_0 = (2.5, 2.0, 0.4, 0.4)^T$. The corresponding initial variance is set as $\mathbf{P}_0 = \text{diag}([1, 1, 0.1, 0.1])$. The variance in the Langevin model is $\mathbf{\Sigma}_k = \text{diag}([1, 1])$. Changing the parameters for the source dynamics, ν and β from Section 2.2, and initialization will lead to different convergence speeds of the algorithm. Other parameters for the tracking algorithms can be found in Table 2. The variance of the measurement noise $\mathbf{R}_k$ for the EKF is set the same as that for the EKPF, which is 1.25e^{-4}, one sample diverging from the measured TDOA. After the EKF step, the samples are relocated around the posterior distribution, and therefore the variance σ_τ^2 in the EKPF is set smaller than it is in the SIR-PF. The parameters q_0 and q_1 depict the effect of reverberation in the PF and EKF step respectively and are chosen heuristically. It is worth pointing out that all parameters in Table 2 are based on extensive experimental studies. For further information, the reader is referred to [42].

4.1. Simulated Room Environment

In simulations, the speed of the source is set at 0.5m/s (1.8km/h), which is one third of a regular pedestrian walking speed, ranging from 5.32km/h to 5.43km/h [51]. Considering that a moving talker within a room is likely to be smooth and slow-paced, this experimental speed is reasonable and com-

parable with the source velocities in [10, 40]. Moreover, 50 frames of speech signal with a length of 128ms are used, at a sampling frequency of 8kHz. Different wall reflection coefficients are set to generate different reverberant environments and different noise conditions are simulated by adding different levels of additive white Gaussian noise (AWGN). The RIR is simulated using the image method [52].

4.1.1. Single Experiment

The tracking results from a single trial under the reverberation time $T_{60} = 0.35$s with wall reflection coefficients of 0.8 are presented. The peaks of the GCC function larger than 0.7 are selected to obtain TDOA measurements. Since the EKF method does not incorporate a reverberant measurement model, it only takes the TDOAs corresponding to the largest peak as its measurements. The TDOA measurements from microphone pair 4 are displayed in Figure 3. It can be observed that the source generated TDOA can be better collected when multiple TDOAs are selected. However, false alarms can arise in TDOA measurements.

The tracking result from a single trial is presented in Figure 4. It shows that the EKF based PF approaches lock on the new source faster than SIR-PF does. Although the EKF is able to find the new source quickly, it is not as robust as the V-EKPF and SIR-PF in dealing with inaccurate TDOA measurements as shown by the track deviations across all time-steps. The MM-EKPF which employs all the multiple TDOA measurements presents the best tracking result. To fully analyze the tracking performance, the RMSE for 100 Monte Carlo (MC) runs is presented in Figure 5. It shows that both the EKF based PF approaches are capable of finding the position of the new

source quickly. However, at some time steps, missing TDOAs can report false alarms, e.g., microphone pair four at time step 71 and thereafter. This leads to a heavy false alarm and miss detection problem. The EKF usually fails to filter the state based on these false measurements and will present unstable results. Subsequently, the following PF cannot draw the samples correctly. This phenomenon can be seen from those peaks in the RMSE made by the EKF and V-EKPF. Since the SIR-PF draws the samples around the previous state estimates, it is not sensitive to the sharp change of source positions and presents the best performance to smooth the inaccurate measurements. However, the drawback is that it cannot lock on the new source quickly. The MM-EKPF is able to cope with false alarms due to reverberation and noise, and also identify the position of the new source quickly.

4.1.2. Different Reverberant and Noisy Environments

The algorithm performance is evaluated in different noisy and reverberant environments. The wall reflection coefficient is set to different values to simulate different environments, and AWGN with different energy levels is added to the received signal to generate different SNRs. For varying reflection coefficients, the SNR is fixed to 30dB, while for varying SNR, the wall reflection coefficient is set to 0.4.

The average RMSE under different reverberant environments over 100 MC runs are presented in Figure 6(a). It shows that the proposed V-EKPF and MM-EKPF approaches are able to track the nonconcurrent multiple sources with good accuracy in the moderate reverberant environment, and perform better than the SIR-PF and EKF. In particular, the MM-EKPF presents the best tracking results since it employs multiple TDOA measure-

ment set as well as the optimum importance function. Figure 6(b) shows the average RMSE in different noisy environments. The performance of the proposed approaches are also better than that of SIR-PF in all experiments. The MM-EKPF consistently presents the best performance in different reverberant and noisy environments. In addition, all the methods are significantly deteriorated in the heavy reverberant and noisy environments.

4.2. Real Room Environment

The performance of the approaches is examined in a real laboratory located at the University of Edinburgh, Scotland. The room has carpet floor, concrete block walls and ceiling, and glass windows covered by hard cardboard with a thickness $\approx$ 0.4cm, as shown in Figure 7. The measured reverberation time is $0.836s$ and the ambient noise level is -40dB [42]. The microphones are mounted on a set of T-bar stands, and the sources are set at a height of 1.33m. The microphone response is omni-directional within the frequency range 0 to 4kHz. The acoustic source used for all recordings is an omnidirectional speaker amounted on a small trolley, as shown in Figure 7. The source is moved via a pulley mechanism and its position is measured using a laser measuring device, by which the sampled locations show that it is moving at a fairly constant velocity. The source signal is taken from the TIMIT database [53]. All measured signals are sampled at $f_s = 44.1$kHz and then downsampled to 8kHz, which is sufficient for ASLT. The frame length is set to 1024 samples, or 128msec, and the source velocity is around 0.5m/s. The TDOA measurements from microphone pair 4 are shown in Figure 8.

Figure 9 presents the tracking results of the various approaches for a single trial, and shows that the MM-EKPF is able to track the sources accurately

and lock on the position of the new source quickly. The EKF fails to do so since large false alarms are presented as the TDOA measurements. Although the SIR-PF is able to track the sources, it cannot catch up with the position of the new source more rapidly than the alternate approaches. Compared to the results from the single experiment of the simulated room environment in Section 4.1.1, the position estimation is degraded since the real room environment is more challenging. Also, uncertainties in the ground truth of the experimental system such as microphone positions and source positions can increase the tracking errors.

Figure 10 gives the tracking results over 100 MC implementations. This statistical result further illustrate that the proposed EKPF algorithms are more accurate than the SIR-PF and EKF in tracking nonconcurrent multiple sources. The RMSE also presents a transition behavior: at the time steps where the source switches, the RMSE increases sharply. The algorithms then converges to the position of the new source. However, the proposed approaches are able to find the position of the new source quickly, while the SIR-PF generally needs much more time steps to lock on the position of the new source. Again, the errors from the experiment system, particular from the estimation of the ground truth of the source positions and microphone positions, can increase the RMSE.

5. Conclusions

This paper addresses a special multiple source tracking case: nonconcurrent multiple acoustic tracking. In such a scenario, one source is active during a period, and then the other follows. Two EKPF approaches are de-

veloped to track the sources, and minimise any lag when the source position changes quickly. The core idea is to utilize an EKF to estimate the state coarsely, and then use a PF to sample around this posterior state estimation, rather than drawn according to the prior information as in the SIR-PF. The information included in the amplitudes of the peaks in the PHAT-GCC that correspond to TDOA measurements have also been incorporated to enhance the performance. Simulated experiments, as well as real recordings, show that the proposed approaches can successfully lock on to the position of the new source more quickly than previous approaches. By incorporating multiple TDOA measurements, the MM-EKPF presents even better performance. While this work assumes that there is one and only one source at each time step, future work includes algorithms to track and detect multiple simultaneously active sources.

References

[1] M. Brandstein, A framework for speech source localization using sensor arrays, Phd Thesis, Brown University, Providence, U.S.A., 1995.

[2] J. H. Dibiase, A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays, Phd Thesis, Brown University, Providence, U.S.A., 2000.

[3] Y. Huang, J. Benesty, G. W. Elko, Passive acoustic source localization for video camera steering, in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process., vol. 2, pp. 909–912, 2000.

[4] R. D. D. Zotkin, L. S. Davis, Multimodal 3-D tracking and event detection via the particle filter, in Proc. IEEE Workshop Detection and Recognition of Events in Video, vol. 2, pp. 20-27 2001.

[5] M. Brandstein, D. Ward, Microphone Arrays. Signal Process. Techniques and Applications, Berlin, Germany: Springer-Verlag, 2001.

[6] F. Talantzis, A. Pnevmatikakis, A. Constantinides, Audio-visual active speaker tracking in cluttered indoors environments, IEEE Trans. Syst., Man, Cybern., B: Cybern., vol. 38, pp. 799-807, 2008.

[7] M. Brandstein, J. Adcock, H. Silverman, A closed-form location estimator for use with room environment microphone arrays, IEEE Trans. Speech and Audio Process., vol. 5, pp. 45-50, 1997.

[8] M. Brandstein, H. Silverman, A practical methodology for speech source localization with microphone arrays, Computer Speech and Language, vol. 11, pp. 91-126, 1997.

[9] S. Gannot, T. G. Dvorkind, Microphone array speaker localizers using spatial-temporal information, EURASIP Journal on Applied Signal Process., pp. 1-17, 2006.

[10] D. Ward, E. Lehmann, R. Williamson, Particle filtering algorithms for tracking an acoustic source in a reverberant environment, IEEE Trans. Speech Audio Process., vol. 11, pp. 826-836, 2003.

[11] J. Benesty, J. Chen, Y. Huang, Time-delay estimation via linear interpolation and cross correlation, IEEE Trans. Speech Audio Process., vol. 12, pp. 509-519, 2004.

[12] T. M. N. Strobel, R. Rabenstein, Speaker localization using steered filtered-and-sum beamformers, in Proc. Erlangen Workshop on Vision, Modeling, and Visualization, vol. 11, pp. 195-202, 1999.

[13] D. B. Ward, R. C. Williamson, Particle filter beamforming for acoustic source localization in a reverberant environment, in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process., vol.2, pp. 1777–1780, 2002.

[14] E. A. Lehmann, A. M. Johansson, Particle filter with integrated voice activity detection for acoustic source tracking, EURASIP Journal on Advances in Signal Process., pp. 1-11, 2007.

[15] E. A. Lehmann, R. C. Williamson, Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments, EURASIP Journal on Applied Signal Process., pp. 1-8, 2007.

[16] M. Fallon, S. Godsill, Multi-target acoustic source tracking using track before detect, in Proc. Workshop on Applications of Signal Process., to Audio and Acoust., vol. 4, pp. 77-80, 2007.

[17] M. Fallon, S. Godsill, Multi-target acoustic source tracking with an unknown and time varying number of targets, in Proc. Joint Workshop on Hands-Free Speech Commun. Microphone Arrays, pp. 77-80, 2008.

[18] M. Fallon, Acoustic source tracking using sequential Monte Carlo, Phd Thesis, Darwin College, University of Cambridge, Cambridge, U.K., 2008.

[19] M. Fallon, S. Godsill, Acoustic source localization and tracking using track before detect, IEEE Trans. Audio, Speech, Lang. Process., vol. 18 pp. 1228-1242, 2010.

[20] M. Fallon, S. Godsill, Acoustic source localization and tracking of a time-varying number of speakers, IEEE Trans. Audio, Speech, Lang. Process., vol. 20, pp. 1409-1415, 2012.

[21] J. Vermaak, A. Blake, Nonlinear filtering for speaker tracking in noisy and reverberant environments, in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process., vol. 5, pp. 3021-3024, 2001.

[22] U. Klee, Tobias, J. McDonough, Kalman filters for time delay of arrival based source localization, EURASIP J. Applied Signal Process., pp. 1-15, 2005.

[23] C. H. Knapp, G. C. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust., Speech and Signal Process., vol. 24, pp. 320-327, 1976.

[24] J. Benesty, Adaptive eigenvalue decomposition algorithm for passive acoustic source localization, J. Acoust. Soc. Amer., vol. 107, pp. 384-391, 2000.

[25] H. Schau, A. Robinson, Passive source localization employing intersecting spherical surfaces from time-of-arrival differences, IEEE Trans. Acoust., Speech and Signal Process., vol. 35, pp. 1223-1225, 1987.

[26] Y. Huang, J. Benesty, G. Elko, R. Mersereati, Real-time passive source

localization: a practical linear-correction least-squares approach, IEEE Trans. Speech Audio Process., vol. 9, pp. 943-956, 2001.

[27] S. Doclo, M. Moonen, Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments, EURASIP Journal on Applied Signal Process., pp. 1110-1124, 2003.

[28] T. G. Dvorkind, S. Gannot, Time difference of arrival estimation of speech source in a noisy and reverberant environment, Signal Process., vol. 85, 177-204, 2005.

[29] J. Chen, J. Benesty, Y. A. Huang, Time delay estimation in room acoustic environments: An overview, EURASIP Journal on Applied Signal Process., pp. 1-19, 2006.

[30] E. A. Lehmann, Particle filtering approach to adaptive time-delay estimation, in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process., vol. 4, pp. 1129-1132, 2006.

[31] Y. T. Chan, K. C. Ho, A simple and efficient estimator for hyperbolic location, IEEE Trans. Signal Process., vol. 42, pp. 1905-1915, 1994.

[32] J. O. Smith, J. S. Abel, Closed-form least-squares source location estimation from range-different measurements, IEEE Trans. Acoust., Speech, and Signal Process., vol. 35, pp. 1661-1669, 1994.

[33] N. Roman, D. Wang, Binaural tracking of multiple moving sources, IEEE Trans. Audio, Speech, and Lang. Process., vol. 16, pp. 728-739, 2008.

[34] M. Raspaud, H. Viste, G. Evangelista, Binaural source localization by joint estimation of ILD and ITD, IEEE Trans. Audio, Speech, and Lang. Process., vol. 18, pp. 68-77, 2010.

[35] J. Vermaak, M. Gangnet, A. Blake, P. Perez, Sequential Monte Carlo fusion of sound and vision for speaker tracking, in Proc. Int. Conf. Computer Vision, pp. 741-746, 2001.

[36] T. Gehrig, K. Nickel, H. Ekenel, U. Klee, J. McDonough, Kalman filters for audio-video source localization, in Proc. IEEE Workshop Application Signal Process., Audio Acoust., pp. 118-121, 2005.

[37] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, I. McCowan, Audio-visual probabilistic tracking of multiple speakers in meetings, IEEE Trans. on Audio, Speech and Lang. Process., vol. 15, pp. 601-616, 2007.

[38] M. Barnard, W. Wang, J. Kittler, S. Naqvi, J. Chambers, A dictionary learning approach to tracking, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process., pp. 981-984, 2012.

[39] D. B. Ward, E. A. Lehmann, R. C. Williamson, Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room, in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process., vol. 5, pp. 117-180, 2003.

[40] W.-K. Ma, B.-N. Vo, S. S. Singh, A. Baddeley, Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach, IEEE Trans. Signal Process., vol. 54, pp. 3291-3304, 2006.

[41] X. Zhong, J. R. Hopgood, Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte Carlo data association, in Proc. IEEE 15th Workshop on Statistical Signal Process., pp. 253-256, 2009.

[42] X. Zhong, A Bayesian framework for multiple acoustic source tracking, PhD. Thesis, The University of Edinburgh, Edinbugh, U.K., 2010.

[43] X. Zhong, J. Hopgood, Nonconcurrent multiple speakers tracking based on extended Kalman particle filter, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process., pp. 293-296, 2008.

[44] A. Levy, S. Gannot, E. A. P. Habets, Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments, IEEE Trans. Audio, Speech and Lang. Process., vol. 19, pp. 1540-1555, 2011.

[45] J. Ianniello, Time delay estimation via cross-correlation in the presence of large estimation errors, IEEE Trans. Acoust., Speech, Signal Process., vol. 30, pp. 998-1003, 1982.

[46] B. Champagne, S. Bedard, A. Stephenne, Performance of time-delay estimation in the presence of room reverberation, IEEE Trans. Speech Audio Process., vol. 4, pp. 148-152, 1996.

[47] J. Chen, J. Benesty, Y. A. Huang, Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments, EURASIP Journal on Applied Signal Process., pp. 2536, 2005.

[48] M. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Signal Process., vol. 50, pp. 174-188, 2002.

[49] D. Simon, Optimal State Estimation, John Wiley and Sons, 2006.

[50] E. A. Lehmann, Particle filtering methods for acoustic source localisation and tracking, PhD Thesis, The Australian National University, 2004.

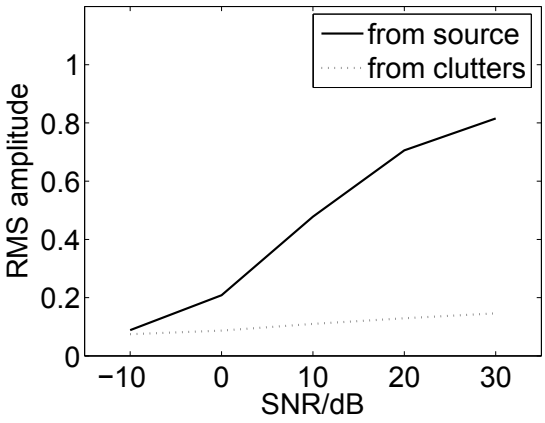
[51] K. Aspelin, Establishing pedestrian walking speeds (Portland State University. 2005. www.westernite.org.).

[52] J. B. Allen, D. Berkley, Image method for efficiently simulating small-room acoust., J. Acoust. Soc. Amer., vol. 65, pp. 943-950, 1979.

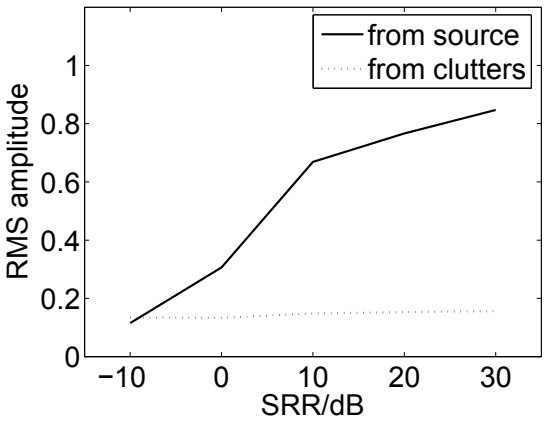
[53] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, V. Zue, TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium, Philadelphia, 1993.

Table 1: Corresponding SRRs generated by different combinations of the source positions and wall reflections.

SRR	-10	0	10	20	30
ρ	0.8	0.8	0.6	0.6	0.1
(x, y)	$(3.6, 3.0)$	$(0.7, 3.0)$	$(0.8, 1.5)$	$(0.8, 1.9)$	$(0.9, 2.1)$



(a) Amplitudes versus SNRs



(b) Amplitudes versus SRRs

Figure 1: The RMS amplitudes of PHAT-GCC TDOA peaks, $\hat{a}_{p,k}^{\ell}$, generated by source and clutter for varying conditions.

Algorithm 1: Proposed EKPF for room acoustic source tracking.

Input: Current TDOA measurements $\mathcal{Z}_k$.

Output: Sources position estimates $\hat{\mathbf{x}}_k$.

Initialisation: draw particles $\mathbf{x}_0^{(i)} \sim \mathcal{N}(\mathbf{x}_0^{(i)}; \mathbf{x}_0, \mathbf{P}_0)$, and set initial weights $\tilde{w}_0^{(i)} = 1/N$.

Over all time steps:

for $k \leftarrow 1$ **to** K **do**

Over all the particles:

for $i \leftarrow 1$ **to** N **do**

- implement EKF to obtain the new samples $\bar{\mathbf{x}}_k^{(i)}$;
- computing the transition density using (40);
- computing the likelihood using (38);
- computing the importance weight using (41);

end

Over all the particles:

for $i \leftarrow 1$ **to** N **do**

- normalize weights: $\tilde{w}_k^{(i)} = w_k^{(i)} / \sum_{i=1}^{N} w_k^{(i)}$;

end

– replicate particles according to weights.

– output the estimates $\hat{\mathbf{x}}_k$.

end

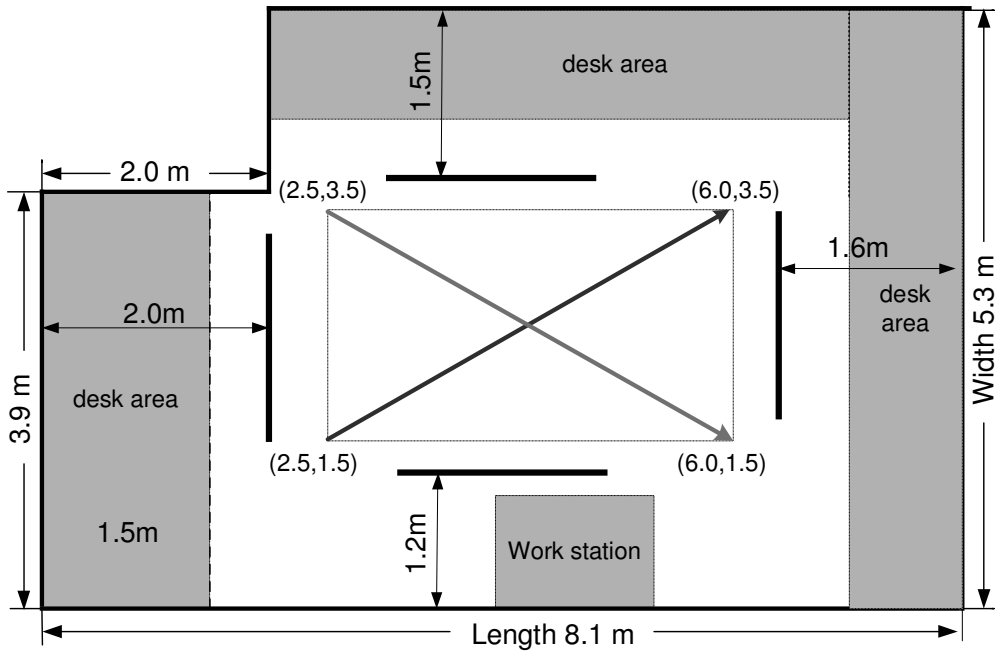


Figure 2: Room environment for experiments. Four microphone arrays equipped with 20 microphones are employed to received the speech signals. The blue and red solid lines represent the source trajectories.

Table 2: Parameter setup for the tracking algorithms.

	$\mathbf{R}_k$	σ_τ^2	q_0	q_1	N
EKF	0.5e^{-4}	-	-	-	-
SIR-PF	-	1.25e^{-4}	0.2	-	500
V-EKPF	1.25e^{-4}	0.5e^{-4}	0.2	0.1	100
MM-EKPF	1.25e^{-4}	0.5e^{-4}	0.2	0.1	100

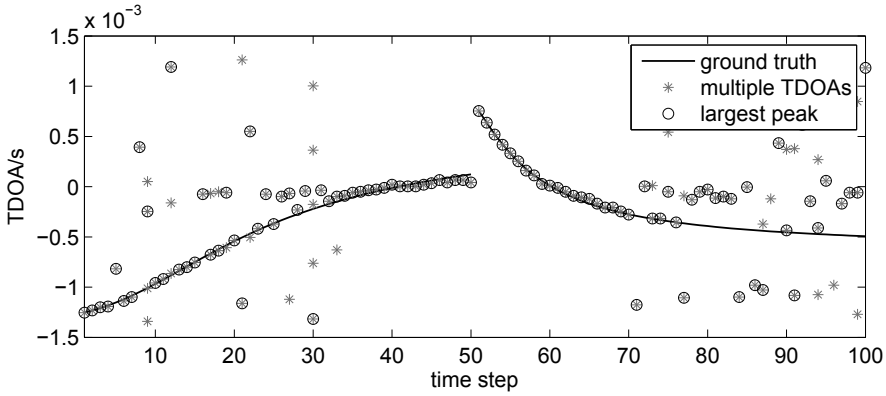
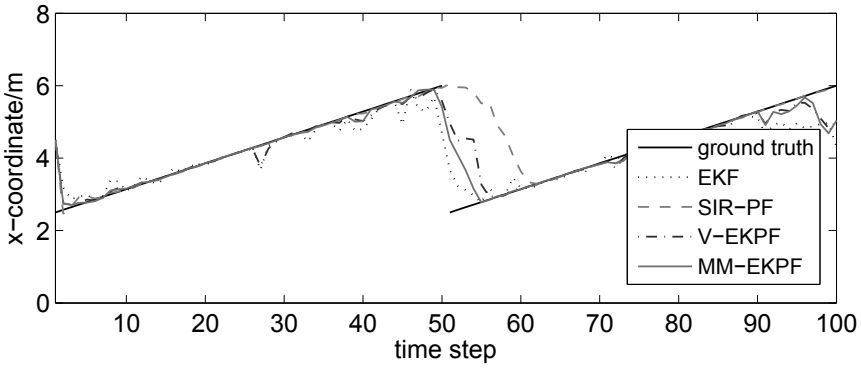
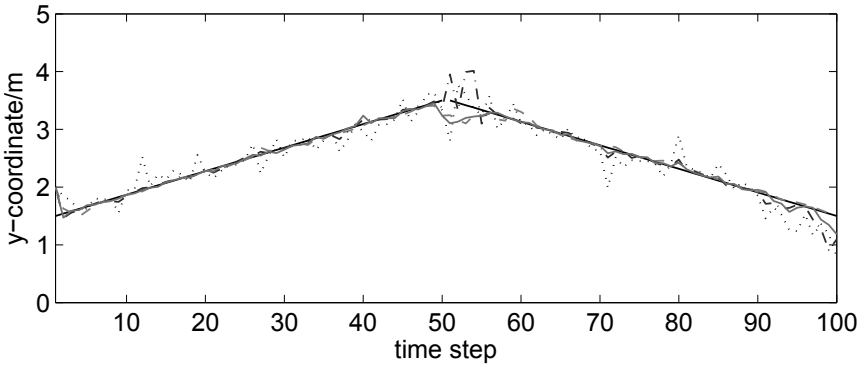


Figure 3: TDOA measurement extracted in a reverberant environment ($T_{60} = 0.35$s) across different microphone pairs.



(a) estimation of x-coordinate



(b) estimation of y-coordinate

Figure 4: Tracking results from a single trial under the reverberant environment (T_{60} = 0.35s): (a) x-coordinate; and (b) y-coordinate.

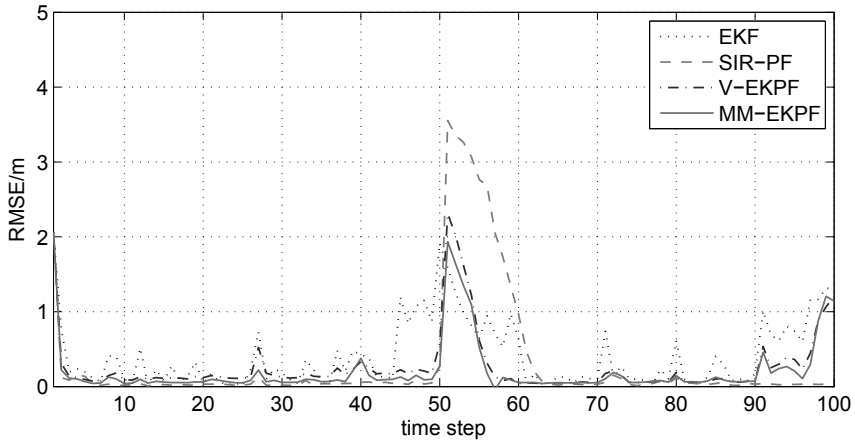
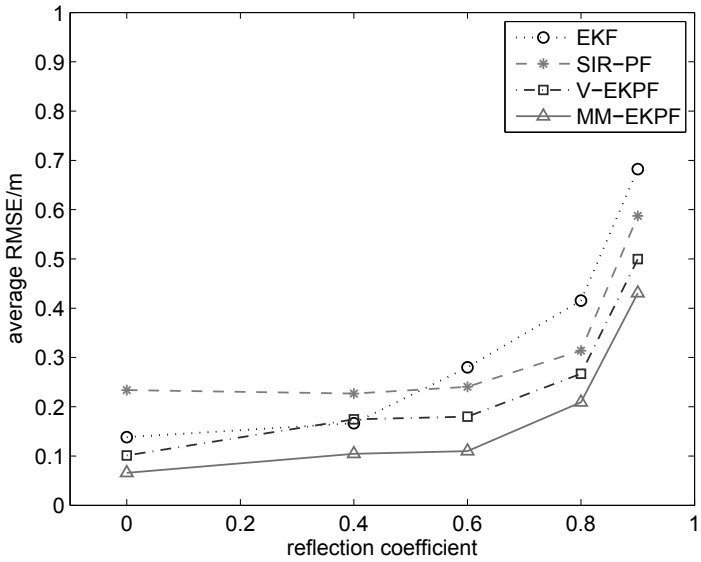
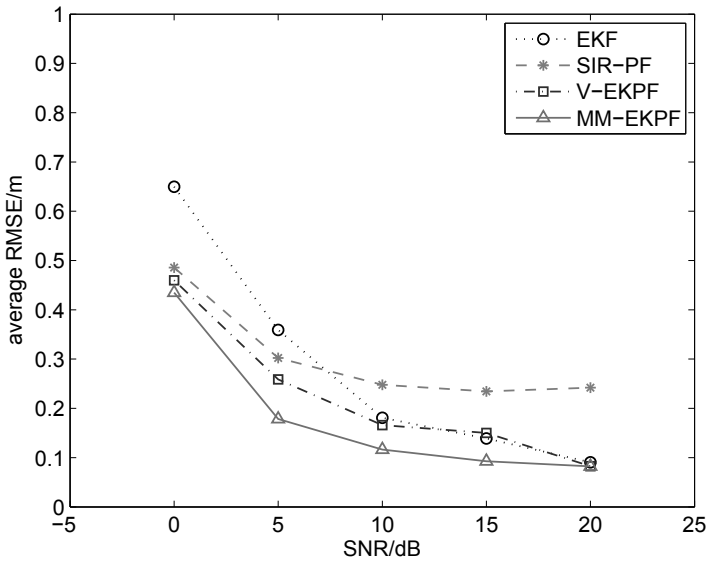


Figure 5: RMSE over 100 Monte Carlo runs under the reverberant environment ($T_{60} = 0.35$s).



(a) Versus different reverberant environments.



(b) Versus different SNRs

Figure 6: Average RMSE under different scenarios

Figure 7: Real audio room environment for performance evaluation. The laboratory is located at the University of Edinburgh, Scotland. The room has carpet floor, concrete block walls and ceiling, and glass windows covered by hard cardboard with a thickness ≈ 0.4cm. The measured reverberation time is $0.836s$ and the ambient noise level is -40dB [42].

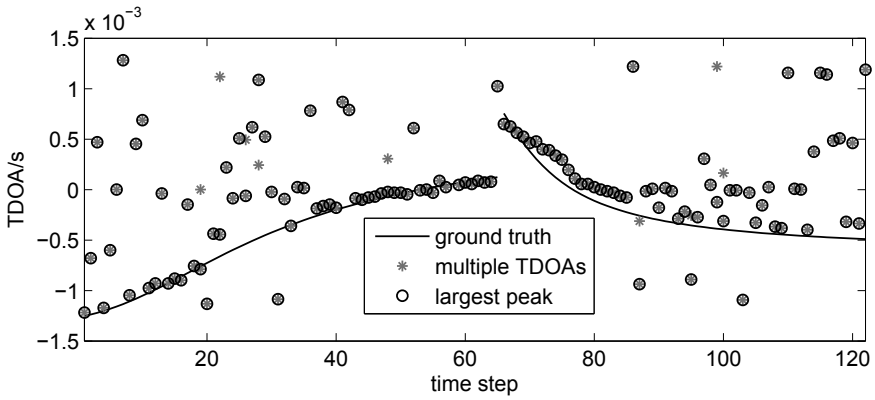
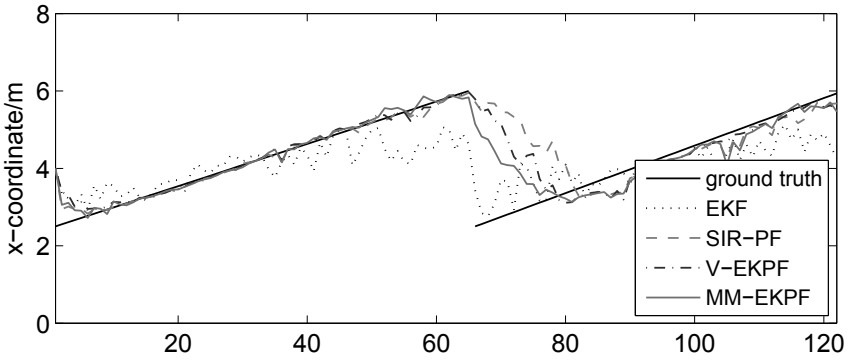
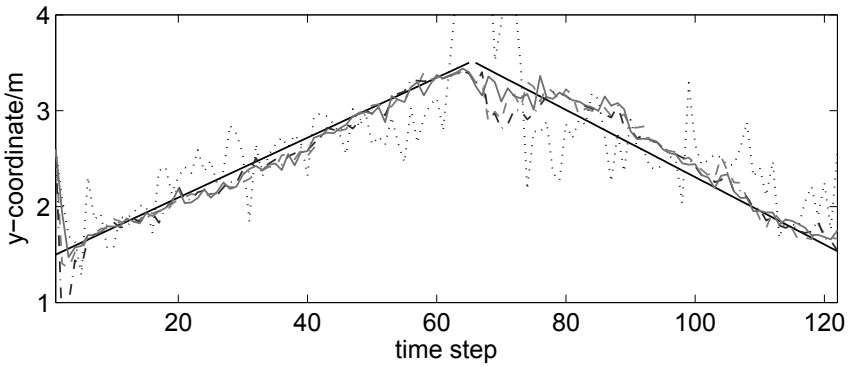


Figure 8: TDOA measurement extracted from microphone pair 4 in a real room environment. Source 1 is active from time step 1 to 65, and then source 2 follows from time step 66 to 123.



(a) estimation of x-coordinate



(b) estimation of y-coordinate

Figure 9: Tracking results from a single trial in the real audio lab environment: (a) x-coordinate; and (b) y-coordinate.

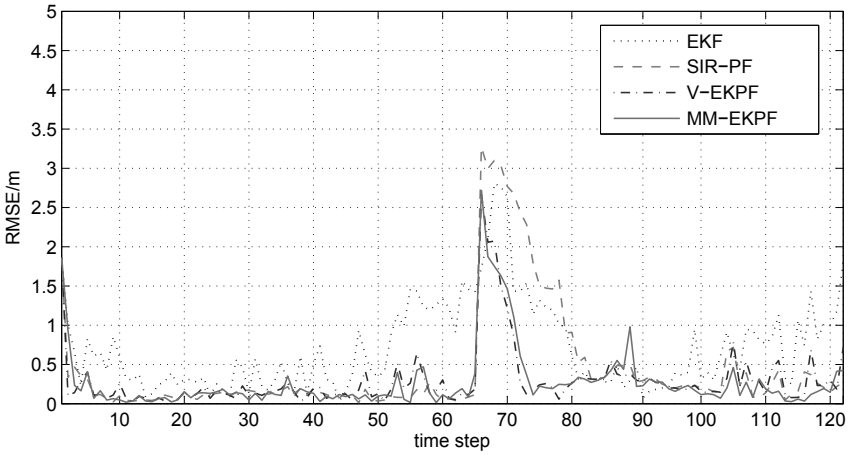


Figure 10: RMSE over 100 Monte Carlo runs in the real audio lab environment.